

The School of Mathematics



THE UNIVERSITY
of EDINBURGH

Understanding the impact of age demographics on hospital admissions in Scotland

by

Christos Giannikos, S2436019

Dissertation Presented for the Degree of
MSc in Statistics with Data Science

July 2023

Supervised by
Dr Vanda Inacio de Carvalho and Prof Simon Wood

Acknowledgments

I am grateful to the supervisors of the project, Dr Vanda Inácio de Carvalho and Prof. Simon Wood, for their constant support and guidance. I would also like to thank Filip Zmuda and Ken Nicholson of Public Health Scotland for providing the data for this project, as well as useful background information.

Word Count

Approximately 5100 words (including the executive summary, main text and references; excluding appendices).

University of Edinburgh – Own Work Declaration

Name: Christos Giannikos

Matriculation Number: S2436019

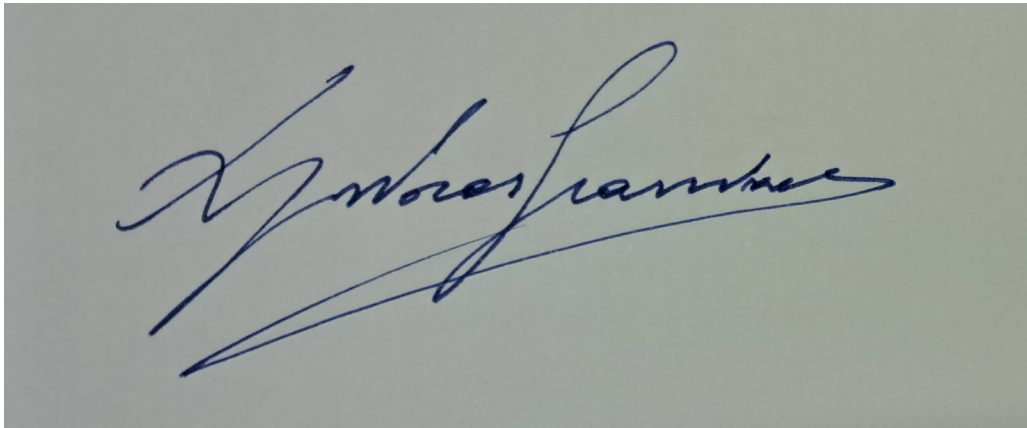
Title of work: Understanding the impact of age demographics on hospital admissions in Scotland

I confirm that all this work is my own except where indicated, and that I have:

- Clearly referenced/listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Not sought or used the help of any external professional academic agencies for the work
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Complied with any other plagiarism criteria specified in the Course handbook

I understand that any false claim for this work will be penalised in accordance with the University regulations (<https://teaching.maths.ed.ac.uk/main/msc-students/msc-programmes/statistics/data-science/assessment/academic-misconduct>).

Signature :

A handwritten signature in blue ink on a grey background. The signature is cursive and appears to read 'Christos Giannikos'.

Date Thursday 29 June 2023

Executive Summary

The significant increase in birth rates, commonly referred to as the "baby boom," which took place from 1946 to 1964, is anticipated to have a considerable impact on the future demand for a wide range of health-care services. As the baby boomer generation continues to age, a substantial portion of the population will be reaching an age where they require more extensive medical attention and specialized services. It is therefore essential to ensure that healthcare systems are equipped to meet the evolving needs of this large population cohort. Towards this end, we use inpatient data from the PHS Open Data website and propose a Negative Binomial Regression model to predict the length of stay of patients in Scottish hospitals over the next 20 years. Furthermore, a Logistic Regression model is applied to assess the probability that a patient's length of stay in a hospital will be greater than zero (counted in days).

The results of our analysis, carried out for Public Health Scotland (PHS), indicate a significant effect on the total length of stay per year. As the baby boomer generation continues to age, we predict a notable increase in the overall length of stay in Scottish hospitals. This projection reflects the rising demand for healthcare services and the urgency of adequately preparing healthcare systems to accommodate the needs of the baby boomer cohort.

Contents

Own Work Declaration	3
Executive Summary	4
1 Introduction	1
2 Literature Review	1
3 Exploratory & Initial Data Analysis	2
3.1 Data Source	2
3.2 Data Cleaning	3
3.3 Data Analysis	3
4 Proposed Methodology	8
4.1 Model Fitting	8
4.2 Results and Predictions	9
4.3 Model Limitations	12
5 Conclusion	13

List of Figures

1	Mean and Median Length of Stay per Age Group	4
2	Total Length of Stay per Year	4
3	Mean and Median Length of Stay by Year	5
4	Mean and Median Length of Stay by Year	6
5	Mean and Median Length of Stay by Location	6
6	Population projections for 2018-2043 in millions	7
7	Total Predicted Length of Stay per Year	10
8	Receiver Operating Characteristic Curve	11
9	11

1 Introduction

The term "baby boomers" refers to the generation of individuals who were born during the post-World War II period, specifically between the years 1946 and 1964. This period is characterized by a significant increase in birth rates and this generation represents a substantial portion of the population in many countries, including Scotland. As baby boomers continue to age, their healthcare needs are gaining increasing significance. The demographic shift resulting from the baby boom has significant implications for the demand and utilization of healthcare services, as this generation moves through different life stages and faces age-related health challenges. By understanding how the length of hospital stays is likely to change as the baby boomer cohort progresses through the healthcare system, policymakers and healthcare providers can prepare for the increased demand for services, in order to effectively respond to the evolving needs of the baby boomer generation. The following analysis is carried out specifically for Public Health Scotland (PHS), an organization responsible for public health services in Scotland. This research is conducted in collaboration with PHS, utilizing their population estimates from 1981 to 2021 and population projections from 2022 to 2043, for the 14 NHS health boards in Scotland, as provided by the PHS Open Data website. The aim of this research is to predict the length of stay for patients in Scottish hospitals for the next 20 years. To achieve this, we will employ a Negative Binomial (Generalized) Regression Model for cases with a length of stay greater than zero (measured in days), while for cases with a length of stay equal to zero, we will implement a Logistic Regression model.

2 Literature Review

Over the years, researchers have employed various methodologies and statistical models to investigate the length of hospital stays. We now give a very brief review of the literature investigating the length of stay for patients in hospitals by employing Negative Binomial models. We consider "Zero-Inflated Negative Binomial Mixed Regression Modeling of Over-Dispersed Count Data with Extra Zeros"[3] by Kelvin K. W. Yau, Kui Wang and Andy H. Lee, published in Biometrical Journal Volume 45, Issue 4, June 2003 (Pages 437–452), who introduce an analytical approach using a zero-inflated negative binomial mixed regression model to examine a dataset of hospital length of stay (LOS) for patients with pancreas disorders. We also consider "Length of hospital stay at Arak (central Iran) maternity clinics using proposed zero-inflated negative binomial modeling" [1] by Rafiei M, Ayatollahi SM, Behboodian J. published in the Pakistan Journal of Biological Sciences :2007, Aug;10(15):2510-2516, where the authors consider a negative binomial model to describe the hospitalisation time of mothers. Additionally, we consider "Factors Associated with Hospital Length of Stay among Cancer Patients with Febrile Neutropenia"[2], by Regis G. Rosa and Luciano Z. Goldani, who used stepwise random-effects negative binomial regression was to identify risk factors for prolonged length of hospital stay for all adult cancer patients with febrile neutropenia at a single tertiary referral hospital in southern Brazil from October 2009 to August 2011.

3 Exploratory & Initial Data Analysis

3.1 Data Source

The primary source of data utilized for this project is the PHS Open Data platform, which consists of a variety of datasets that can offer insights into various aspects of healthcare services. We primarily focus our interest on the dataset called "Activity by Board of Treatment, Age and Sex," which contains information related to hospital admissions and patient demographics, with a breakdown by age and sex for both inpatient and day case activities. The data within this dataset have been aggregated into monthly quarters, providing a granular view of healthcare activities over time. Additionally, the dataset includes variables such as location, health board and admission type, allowing for a comprehensive analysis of patient characteristics and their association with hospital admissions. In greater detail, the variables (columns) present in the dataset are the following:

- Quarter: The reference period represented as a quarter (in the format YYYYQN).
- HB: A 9-digit code representing the health board of treatment based on the boundaries as of 1st April 2019. It includes a country code for Scotland and 6-digit codes for special boards.
- Location: A 5-digit code for the hospital, a 9-digit code for the health board of treatment, and a country code for Scotland. It also includes 6-digit codes for special boards.
- AdmissionType: The admission type of the patient, categorized as Elective Inpatients, Emergency Inpatients, Transfers, All Daycases, All Inpatients, or All Inpatients and Day Cases.
- Sex: The sex of the patient, classified as Female or Male.
- Age: The age of the patient, categorized into 10-year age bands.
- Episodes: The number of episodes, representing the frequency of hospital admissions for a particular combination of variables.
- LengthOfEpisode: The total length of episodes, indicating the cumulative duration of hospital stays for a specific combination of variables.
- AverageLengthOfEpisode: The average length of an episode, calculated as the total length of episodes divided by the number of episodes.
- Stays: The number of stays, representing the count of hospital stays for a given combination of variables.
- LengthOfStay: The total length of stays, indicating the cumulative duration of hospital stays for a specific combination of variables.
- AverageLengthOfStay: The average length of a stay, calculated as the total length of stays divided by the number of stays.

Therefore, each entry (row) of the dataset represents the data related to all patients of the same sex and age category regarding a specific hospital over the same reference period. Furthermore, within the dataset, there are statistical qualifiers associated with certain columns, including Quarter, HB (Health Board), Location, AdmissionType, AverageLengthOfEpisode, and AverageLengthOfStay. These qualifiers serve to provide additional information and context for the respective data points, i.e. whether the data is provisional, aggregated or if the specific data item is not applicable, while they also help ensure accurate analysis and interpretation by indicating the nature of the data. It must be noted that, according to the definitions provided at the PHS website, an episode is defined as an event where a patient gets admitted in a hospital under a specialty. A move to another specialty initiates a new episode. (e.g. a new diagnosis). Note that Inpatient is every patient which occupies a staffed bed in a hospital and either remains overnight or is expected to remain overnight but is discharged earlier. An Inpatient's admission can be an Emergency, an Elective or a Transfer. On the contrary,

a Day Case is when a patient makes a planned attendance for one day to a specialty for clinical care and requires the use of a bed or a trolley. Even though a Day Case is usually completed on the same day, the patient may need to be admitted as an Inpatient, if they are not able to be discharged.

3.2 Data Cleaning

As part of the data cleaning process, it is important to ensure the accuracy and reliability of the data. Therefore, in our analysis we decided to focus only on the entries that pertain to inpatient and day case admissions, excluding transfers. This allows us to concentrate on the specific types of hospital admissions that are relevant to our research. Additionally, we identified certain Health Boards that needed to be excluded from our analysis for various reasons. These Health Boards are as follows:

1. Health Board S27000001: This entry represents a non-NHS provided location, which is not within the scope of our analysis. Therefore, we choose to exclude it from our dataset.
2. Health Board S27000002: This entry is marked as "not applicable," indicating that it does not provide relevant information for our analysis.
3. Health Board S92000003: This entry corresponds to the entire country of Scotland. Since our analysis focuses on specific Health Boards within Scotland, we exclude this entry from our dataset to avoid duplicate values.
4. Health Board SB0811: This entry represents a special Health Board for which we do not have a population estimate.
5. Health Board SN0811: This entry denotes a closed Health Board, which is no longer operational. As it is not relevant to our analysis, we remove it from our dataset.

By excluding these specific entries and Health Boards from our dataset, we ensure that our analysis is based on accurate and relevant data, thereby enhancing the reliability of our findings. In addition, to ensure data accuracy and reliability, we need to exclude certain cases that appear to be erroneous. These cases include:

1. Entries where the number of episodes is 0, but the length of stay is greater than 0. This scenario is logically inconsistent since a length of stay cannot exist without any recorded episodes.
2. Entries where the number of stays is 0, but the length of stay is greater than 0. Similar to the previous case, this situation indicates an inconsistency in the data.
3. Cases where the health board of treatment is the same as the provided location. This suggests that the data may have been derived from the respective health board total, leading to duplication.
4. Cases where the average length of stay has the qualifier "z," indicating that it is not applicable and the actual data are missing.

By excluding these erroneous cases from the analysis, we can ensure the accuracy of the dataset, leading to more reliable insights and conclusions.

3.3 Data Analysis

We proceed in our data analysis by examining the relationship between age and the length of stay for the whole of Scotland between 2017 and 2022. Figure 1 presents the mean and median length of stay values across the different age groups.

Upon analyzing the data, a distinct pattern in the mean and median length of stay values is observed as age progresses. The mean length of stay initially increases with age until reaching a peak around 80 years, after which it starts to decrease. This suggests that, on average, older individuals

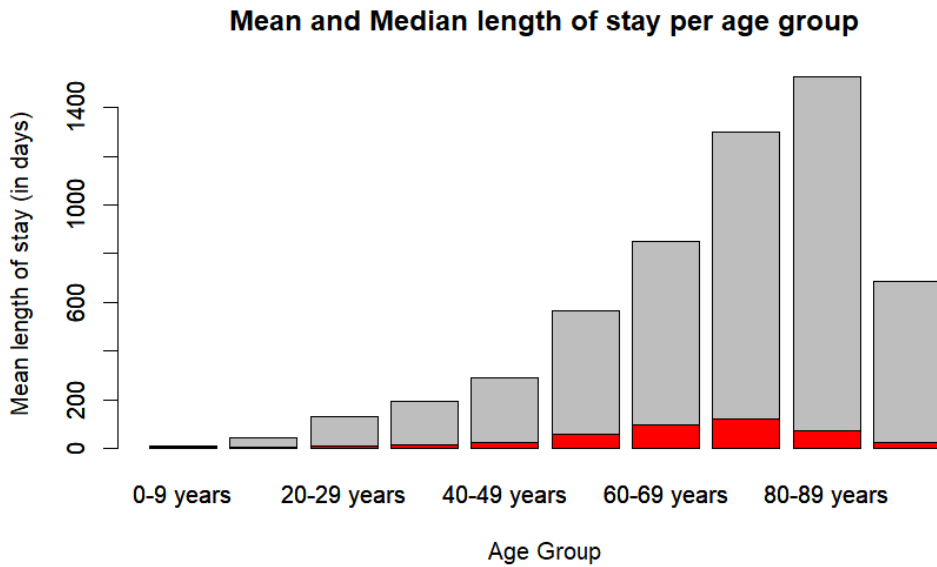


Figure 1: Mean and Median Length of Stay per Age Group

tend to have longer hospital stays until a certain point, after which the length of stay begins to decline. In contrast, the median length of stay follows a slightly different trend. It increases steadily with age until around 70 years, indicating that the majority of patients within each age group have relatively consistent lengths of stay. However, after 80 years, the median length of stay starts to decrease. This could be due to various factors, including changes in healthcare practices, improved medical interventions, or an increased focus on early discharge and care in non-hospital settings for older adults. This discrepancy between the mean and median values suggests that there might be some extreme or prolonged hospital stays among a small portion of the population, which impacts the mean value more significantly. On the other hand, the median provides a better representation of the typical length of stay experienced by the majority of individuals within each age group.

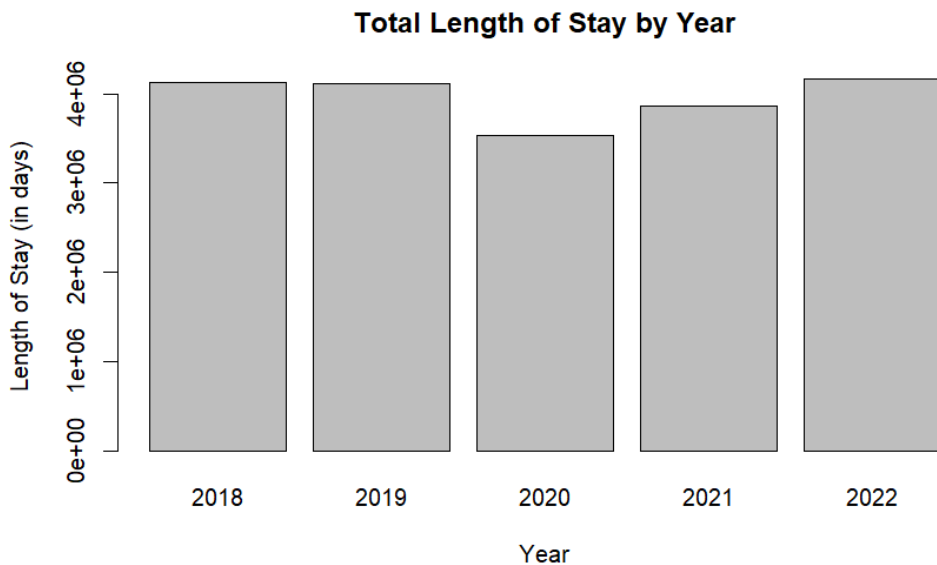


Figure 2: Total Length of Stay per Year

We further explored the data by analyzing the total length of stay for all hospitals across Scotland over the last five years. Figure 2 presents the total length of stay by year. In 2018 and 2019, the

total length of stay remained relatively stable, with a value of around 4,100,000 days. This indicates a consistent level of healthcare utilization and patient stays during those years. However, a significant drop in the total length of stay was observed in 2020, with the value decreasing to approximately 3,500,000 days. This sudden decrease can be attributed to the unprecedented COVID-19 pandemic, which had a profound impact on healthcare systems worldwide. The implementation of lockdown measures, reduced elective procedures, and changes in healthcare-seeking behavior likely contributed to the decline in the total length of stay during this period. As the healthcare system began to adapt to the challenges posed by the pandemic, we observed a slight recovery in the total length of stay in 2021 and 2022. The values increased to 3,800,000 and 4,100,000 days, respectively. This indicates a gradual rebound in healthcare utilization and patient stays as the effects of the pandemic were managed and healthcare services resumed closer to pre-pandemic levels.

We also examined the mean and median length of stay per year for all hospitals across Scotland. Figure 3 presents the mean and median length of stay by year.

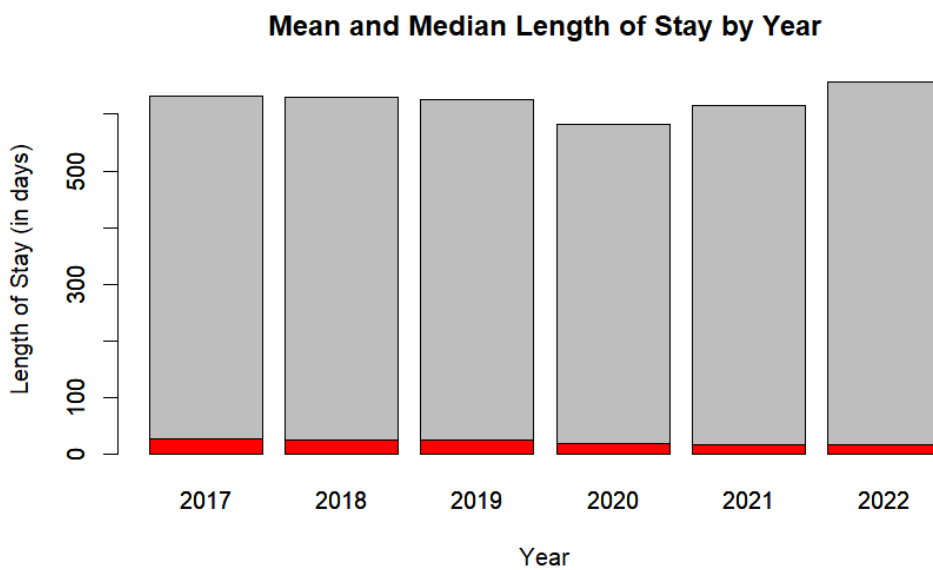


Figure 3: Mean and Median Length of Stay by Year

Looking at the mean length of stay, we observed relatively consistent values from 2017 to 2019, with a slight fluctuation around 630 days. However, in 2020, there was a notable decrease in the mean length of stay to almost 580 days. This decline can be attributed to various factors, including the COVID pandemic and subsequent changes in healthcare practices, increased efficiency in patient management, or shifts in patient demographics. In the subsequent years, 2021 and 2022, the mean length of stay increased again, reaching approximately 615 and 657 days, respectively. This indicates a potential reversal of the decreasing trend observed in 2020 and suggests a rebound in the average length of hospital stays.

When examining the median length of stay, we observed a relatively stable pattern over the years. From 2017 to 2019, the median length of stay remained consistent, ranging from 26 to 24 days. In 2020, there was a slight decrease in the median length of stay to 19.5 days. This decreasing trend continued in 2021 and 2022, where the median length of stay slightly decreased to 16 and 17 days, respectively. These findings suggest that while the mean length of stay fluctuated over the years, the median length of stay remained relatively stable. This could be explained by the fact that the median provides a measure of the central tendency that is less affected by extreme values, making it a robust indicator of the typical length of hospital stays.

Additionally, we also conducted an analysis of the length of stay by admission type, specifically for "All Day cases," "Elective Inpatients," and "Emergency Inpatients." The findings are summarized in Figure 4.

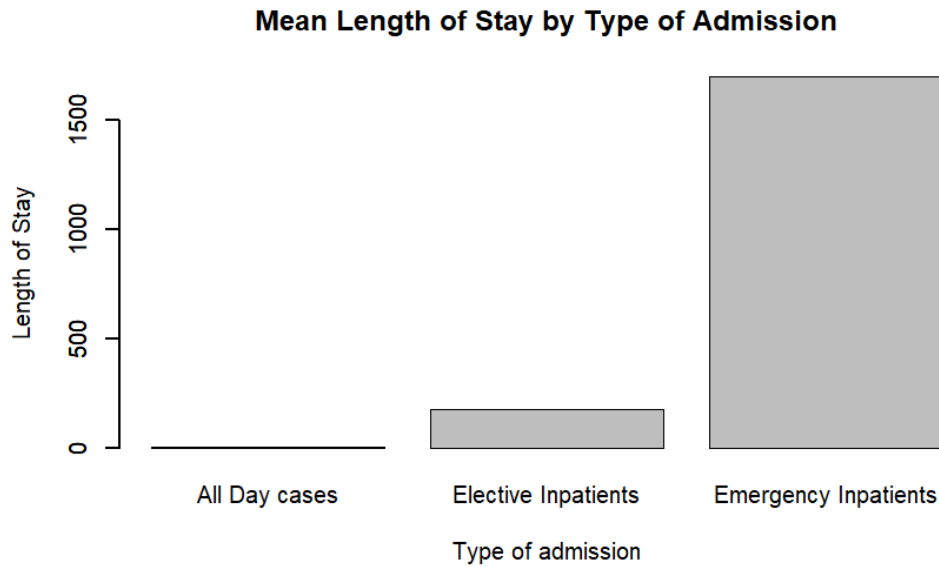


Figure 4: Mean and Median Length of Stay by Year

For day cases, the average length of stay is approximately 5 days, indicating relatively short durations. This is to be expected, since day cases typically involve patients who receive clinical care and require the use of a bed or a trolley but are discharged on the same day. In contrast, elective inpatients, who undergo planned procedures or treatments, have an average length of stay of around 175 days. These longer durations can be attributed to the nature of elective admissions, which often involve more complex medical interventions or surgeries requiring extended recovery periods. For emergency inpatients, the average length of stay is approximately 1699.42 days. These patients require urgent medical care or treatment for critical conditions, resulting in significantly longer hospital stays. The severity and complexity of these cases contribute to the extended durations of their hospitalization. These findings highlight the varying lengths of stay based on the admission type. Day cases have the shortest stays, elective inpatients have longer stays, and emergency inpatients experience the longest durations in the hospital.

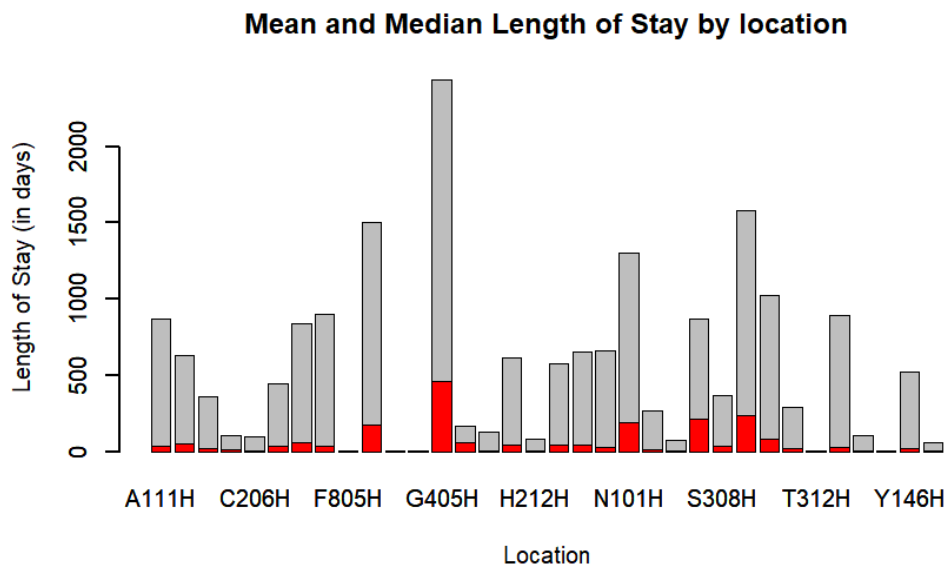


Figure 5: Mean and Median Length of Stay by Location

In addition, we investigate the mean and median length of stay for different locations. These are demonstrated in Figure 5. The plot displays the mean and median length of stay for different locations. Each bar represents a location, and the height of the bar corresponds to the mean length of stay for that location. The red line within each bar represents the median length of stay. It is evident that the mean length of stay varies across different locations. Locations "G107H" and "T101H", corresponding to the Glasgow Royal Infirmary and the Ninewells Hospital in Dundee respectively, have the highest mean length of stay, exceeding 1500 days, while locations "C121H" and "C206H", corresponding to the Lorn and Islands Hospital and the Vale of Leven General Hospital, have relatively low mean lengths of stay, below 150 days. On the other hand, the median length of stay provides information about the central tendency of the data. Locations "G107H" and "V217H" (Fort Valley Royal Hospital) have the highest median length of stay, around 180 days, while locations "C206H" and "F805H" (Victoria Maternity Unit) have the lowest median lengths of stay, around 5 days. It should also be noted that there is considerable variation in length of stay among different locations, as indicated by the range between the highest and lowest mean and median values. This suggests that the length of stay is influenced by factors associated with specific locations, since these might represent different types of healthcare facility and different types of demographic composition in their catchment areas.

Finally, we take a look at the population projections for 2018-2043, as demonstrated in Figure 6.

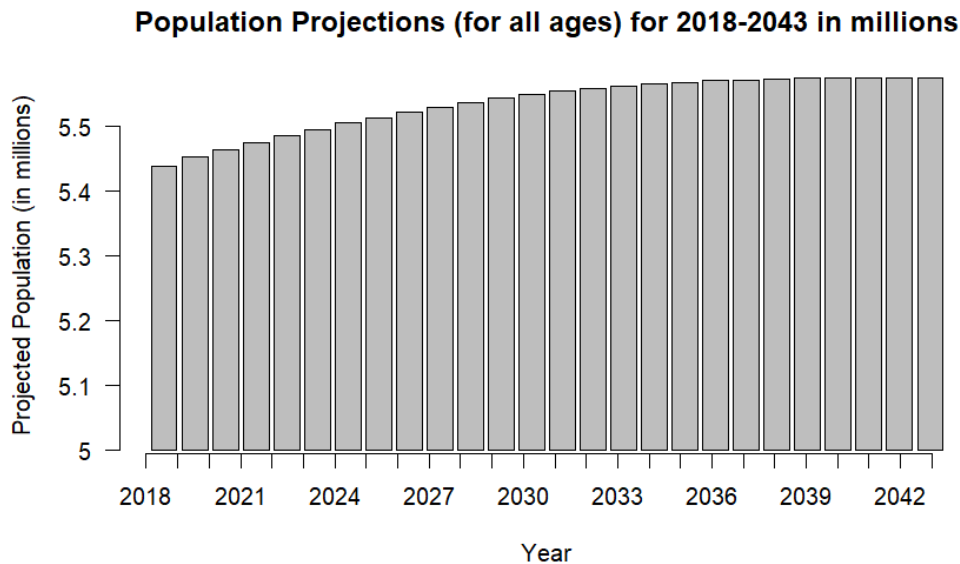


Figure 6: Population projections for 2018-2043 in millions

The population projections show a modest upward trend, indicating an overall increase in population over the years, suggesting a small but stable and sustained growth rate over the specified time period.

4 Proposed Methodology

4.1 Model Fitting

The distinct characteristics and handling of Day Cases and Inpatients in the healthcare system provide a logical rationale for employing two different models, one for each of them. Given that the data consist of count data representing the length of stay in days, it is reasonable to consider a Negative Binomial Generalized Linear Model (GLM) for cases with a length of stay greater than zero. Additionally, for cases with a length of stay equal to zero, a logistic regression model is appropriate to model the probability of same-day discharges. In the case of the factor variables Sex, Age, Location and AdmissionType, we will use dummy coding or indicator variables to represent the different categories. Sex has two levels (Male and Female), Age has 10 levels (0-9 years old, 10-19 years old,...,90 plus years old), Location has 34 levels (location A210H,..., Z102H), and Admission Type has 3 levels (Day Case, Emergency Inpatient and Elective Inpatient).

Before we proceed to describing the mathematical modelling, it must be noted that, since population projections are only available for Health Boards, for each location we consider the population of the corresponding Health Board. Additionally, we split the Quarter variable, which is categorical, to two integer variables, $Year$ and $Quarter \in 1, 2, 3, 4$. Therefore we proceed by proposing the following two models:

1. The Negative Binomial GLM is commonly used to model count data with overdispersion, which occurs when the variance of the data exceeds the mean. It extends the traditional Poisson regression model by introducing an additional parameter to account for the extra dispersion. The model assumes that (for the cases where it's strictly larger than 0) the Length of Stay follows a Negative Binomial distribution and its logarithm is connected with the predictors through a linear relationship. In other words, the model can be represented as:

$$LengthOfStay \sim NegativeBinomial(r, p)$$

$$\mu = E[LengthOfStay]$$

and

$$\begin{aligned} \log(\mu) = & \beta_0 + \beta_1 Sex_{Female} + \beta_2 Age_{10-19yearsold} + \dots + \beta_{10} Age_{90plus} + \\ & + \beta_{11} Location_{A210H} + \dots + \beta_{44} Population + \beta_{45} Quarter + \\ & + \beta_{46} AdmissionType_{ElectiveInpatients} + \beta_{47} AdmissionType_{EmergencyInpatients} \end{aligned}$$

where:

- $r \geq 0$ and $p \in [0,1]$ are the distribution parameters
 - $\log(\mu)$ represents the logarithm of the mean count of length of stay.
 - $Sex_{Female}, Age_{10-19yearsold}, \dots, Location_{A210H}, \dots, AdmissionType_{ElectiveInpatients}$ and $AdmissionType_{EmergencyInpatients}$ are indicator variables that take the value 1 if the aggregated cases belong to that category and 0 otherwise.
 - β_0, β_1, \dots are the regression coefficients for the corresponding predictors.
2. Logistic regression is a widely used model for binary outcomes. It models the log-odds of the probability of an event occurring. Consequently, in our case, a logistic regression model can be used to estimate the probability of a case having a length of stay equal to zero (same-day discharges). By implementing this model we take into account the data points that we previously discarded for our Negative Binomial model. The model equation can be written as:

$$Y = \begin{cases} 1, & \text{if LengthOfStay} > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$Y \sim \text{Bernoulli}(p), E[Y] = p$$

$$\begin{aligned} \log\left(\frac{p}{1-p}\right) = & \gamma_0 + \gamma_1 \text{Sex}_{Female} + \gamma_2 \text{Age}_{10-19yearsold} + \dots + \gamma_{10} \text{Age}_{90plus} + \\ & + \gamma_{11} \text{Location}_{A210H} + \dots + \gamma_{44} \text{Population} + \gamma_{45} \text{Quarter} + \\ & + \gamma_{46} \text{AdmissionType}_{ElectiveInpatients} + \gamma_{47} \text{AdmissionType}_{EmergencyInpatients} \end{aligned}$$

where:

- Y is a binary variable that
- p represents the probability of the length of stay being equal to zero
- $\log(\frac{p}{1-p})$ represents the logarithm of the odds ratio of same-day discharges
- $\gamma_0, \gamma_1, \dots$ are the regression coefficients for the corresponding predictors that represent the log-odds of the length of stay being equal to zero for each category of the factor variables, controlling for the other predictor variables. These coefficients are interpreted as the change in log-odds of the outcome for a one-unit change in the corresponding predictor, holding other predictors constant.

4.2 Results and Predictions

After fitting the Negative Binomial (Generalized) Regression Model in the given data, the following results can be observed:

1. The intercept term in the model is statistically significant ($p < 2e^{-16}$), indicating that when all other predictor variables are held constant, there is a negative log-rate of change in the expected length of stay.
2. Among the factor variables, Sex, Age, Location, and AdmissionType all show significant associations with the expected length of stay. For example, being male (SexMale) is associated with a decrease in the expected length of stay compared to being female ($p < 2e-16$).
3. Similarly, different age categories (e.g., Age10-19 years, Age20-29 years, etc.) show varying effects on the expected length of stay compared to a reference category (e.g., Age0-9 years).
4. The location of the hospital also has a significant impact on the expected length of stay, with various locations (e.g., LocationB120H, LocationC121H, etc.) showing different effects compared to a reference location (e.g., LocationA210H).
5. The type of admission (AdmissionType) is strongly associated with the expected length of stay, with elective inpatients (AdmissionTypeElective Inpatients) and emergency inpatients (AdmissionTypeEmergency Inpatients) having significantly longer expected stays compared to the reference category.
6. The dispersion parameter (r) of the negative binomial distribution is estimated to be 2.0760, and the standard error of the dispersion parameter is 0.0200.
7. The model's goodness of fit can be evaluated based on the deviance statistics. The null deviance, which represents the deviance when only the intercept is included in the model, is 143518. The residual deviance, which represents the deviance after including the predictor variables, is 22155 and the Akaike Information Criterion (AIC) value is 261286.

In our analysis of the relationship between age and length of stay in hospitals in Scotland, we initially used population estimates provided by PHS Open Data. However, in order to make future predictions, we now incorporate the respective population projections for the upcoming years. The actual length of stay from 2018 to 2022 and the predicted total length of stay per year from 2023 to 2043 is presented in Figure 5 below:

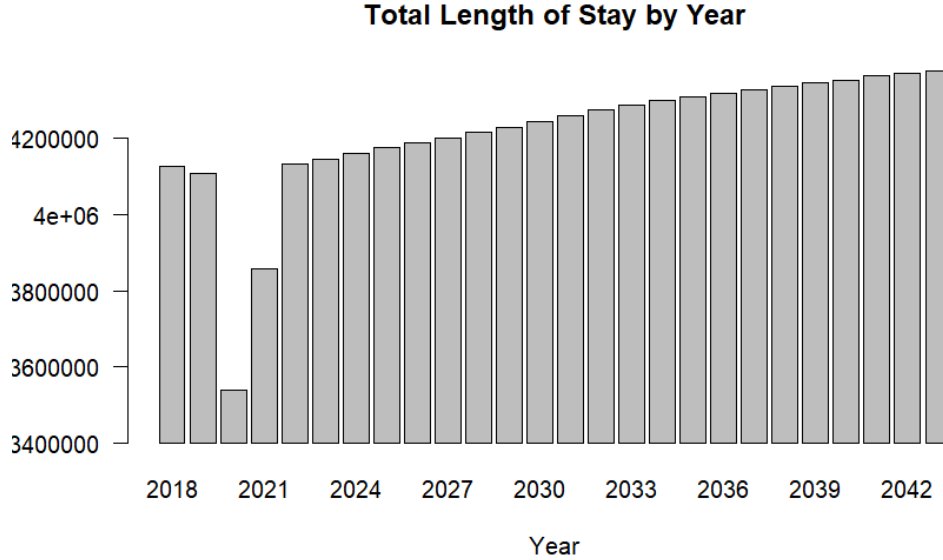


Figure 7: Total Predicted Length of Stay per Year

As previously noted, the total length of stay exhibited a significant decline in 2020, primarily attributed to the profound impact of the COVID-19 pandemic on healthcare systems worldwide. However, based on our population projections and analysis, we anticipate a gradual recovery and adaptation of the healthcare system in the coming years, leading to an expected rise in the total length of stay. According to our projections, in 2023, the total length of stay is predicted to reach approximately 4,150,000 days, indicating a noteworthy rebound in healthcare utilization and patient stays compared to the low observed in 2020. This resurgence reflects the ongoing efforts to restore and optimize healthcare services, as well as the increasing demand for medical care as individuals seek necessary treatments and procedures. Looking further into the future, our model suggests a relatively stable pattern in the total length of stay per year, with minor fluctuations and slight increases anticipated in certain years. By the year 2043, the estimated total length of stay is projected to reach approximately 4,400,000 days, signifying a modest yet consistent upward trend over the forecast period.

In addition, after fitting the Logistic Regression model in our data, we obtain the evaluation metrics of the model, which suggest high performance in predicting the non-zero length of stay.

- The sensitivity, also known as the true positive rate, indicates that the model correctly identified 99.6% of cases where the length of stay was non-zero.
- The false positive rate, represented by $1 - \text{specificity}$, shows that the model incorrectly classified 4.3% of cases where the length of stay was actually zero.
- The overall accuracy of the model is 98.1%, indicating a high level of correct predictions.
- The precision, which measures the proportion of correctly predicted non-zero length of stay cases out of all predicted non-zero cases, is 97.3%. This suggests that when the model predicts a non-zero length of stay, it is correct approximately 97.3% of the time.
- The F1 score, a harmonic mean of precision and sensitivity, is 98.5%, indicating a good balance between precision and sensitivity.

- The receiver operating characteristic (ROC) curve, generated using the predicted probabilities from the model, illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity). As demonstrated in Figure 8, the area under the ROC curve is 0.9962, indicating excellent discriminative power of the model.

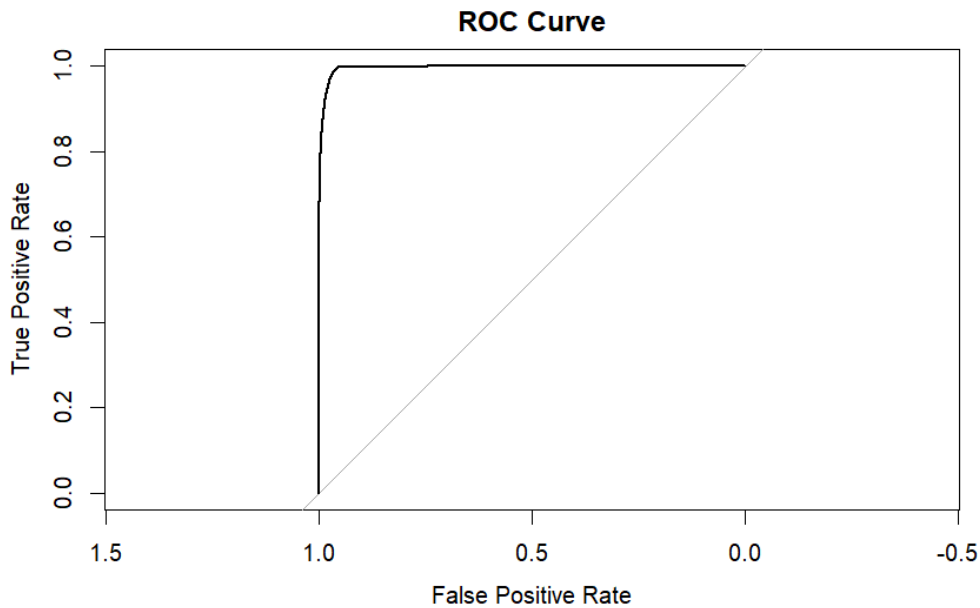


Figure 8: Receiver Operating Characteristic Curve

After fitting the Logistic regression model to the data, we observe that the trend of the proportions of yearly cases that are predicted to have a length of stay greater than 0, as demonstrated in Figure 9 shows a gradual decrease over time. In particular, in the year 2018, the proportion of positive predictions is 0.640, indicating that around 64% of the cases are predicted as positive. As the years progress, the proportion gradually declines, reaching 0.266 in the year 2043, indicating that only around 27% of the cases are predicted as positive.

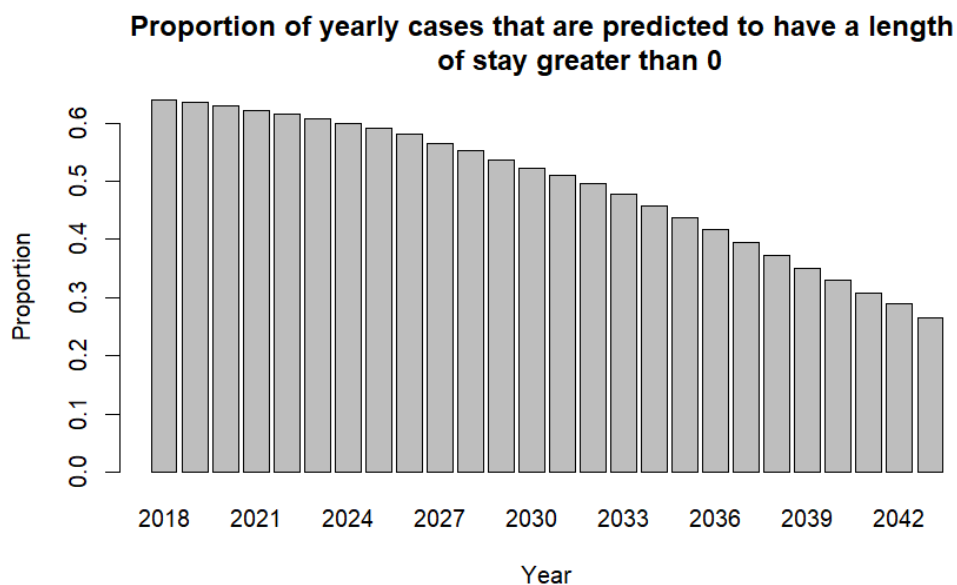


Figure 9:

This decreasing trend suggests that, according to the logistic regression model, the likelihood of a positive prediction decreases as the years go by. It implies that there might be factors or changes over time that contribute to a lower probability of the outcome being positive.

4.3 Model Limitations

While our analysis provides useful insights into the length of stay for patients in Scottish hospitals and offers predictions for future trends, it is important to acknowledge the limitations of our models and methodology. The following limitations should be considered:

- The population projections utilized for predictions are subject to change, since they are based on current demographic trends and assumptions about future population growth. Therefore, they do not account for unforeseen events, such as policy changes, technological advancements or changes in government policies that can significantly impact population dynamics. As a result, the proposed models cannot account for unpredictable factors that may impact the length of stay in the future.
- The analysis conducted focuses on Scottish hospitals and the baby boomer generation. The findings may not directly apply to other regions or age groups, as healthcare systems, demographics, and patient characteristics can vary significantly.
- The models used in the analysis establish associations between variables but do not necessarily indicate causality. This implies that the predicted increase in the total length of stay may not necessarily be due to the ageing of the boomer cohort. This is just one possible factor among many that could influence the length of hospital stays. Other variables, such as changes in healthcare policies, advancements in medical treatments, shifts in patient demographics, or variations in hospital practices, may also play a role. Additionally, there may be interactions and complex relationships between different factors that affect the total length of stay.

5 Conclusion

This report aims to analyze the length of stay for patients in Scottish hospitals and predict future trends using a Negative Binomial Regression model and Logistic Regression model. The study focused on the baby boomer generation, a significant portion of the population that is aging and will require increased healthcare services in the coming years.

Through exploratory data analysis, we identified patterns in the relationship between age and length of stay. The mean length of stay initially increases with age, peaking around 80 years. On the other hand, the median length of stay showed a steady increase with age until around 70 years, followed by a decrease. These findings suggest that while older individuals tend to have longer hospital stays on average, the majority of patients within each age group experience relatively consistent lengths of stay.

Furthermore, our analysis revealed the impact of the COVID-19 pandemic on healthcare utilization and patient stays. The total length of stay in Scottish hospitals decreased in 2020 due to pandemic-related factors such as lockdown measures and reduced elective procedures. However, there was a gradual recovery in the following years, indicating a return to pre-pandemic levels of healthcare utilization.

Moving forward, the models predict an increase of 6% in the total length of stay between 2023 and 2043, while during the same period the population is expected to grow only by 1.4%. These figures indicate that future demand for hospital services is expected to rise steadily as time progresses and at a higher rate than population growth. By anticipating the needs of the aging baby boomer generation, healthcare systems can prepare for the evolving requirements, such as specialized services and longer lengths of stay.

However, it is important to note that the analysis is subject to limitations such as the potential for changes in population projections, making them less reliable for long-term predictions. The findings may not be applicable to other regions or age groups due to varying healthcare systems and demographics. Finally, the models establish associations between variables but do not determine causality, suggesting that factors other than the ageing boomer cohort may influence the increase in total length of stay.

References

- [1] M. Rafei, J. Behboodian, and S. Ayatollahi. "length of hospital stay at arak (central iran) maternity clinics using proposed zero-inflated negative binomial modeling. *Pakistan Journal of Biological Sciences*, 45(2):2510–2516, 2007.
- [2] R. G. Rosa and L. Z. Goldani. Factors associated with hospital length of stay among cancer patients with febrile neutropenia. *Pakistan Journal of Biological Sciences*, 45(3), 2014.
- [3] K. Yau, K. Wang, and A. Lee. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal Volume*, 45(1):437–452, 2003.